Automated Fake News Detection Using Machine Learning and BERT

Business Problem

Purpose:
The proliferation of fake news presents major challenges for society, influencing public opinion, elections, and health decisions. With the scale and speed of misinformation online, manual detection is infeasible. This project demonstrates an automated, explainable system for distinguishing fake from real news using classical machine learning and deep learning models.

Background / History

Context:
Fake news has historically existed but has surged in influence with digital and social media. Traditional manual fact-checking cannot keep pace. Machine learning models, capable of capturing language patterns and content cues, are increasingly used for large-scale detection of misinformation.

Data Explanation

Data Sources

- **Kaggle Fake and Real News Datasets**:
  Contains ~40,000 articles with "fake" or "real" labels. Fields: title, text, subject, date.

Data Preparation Steps

- Null and duplicate removal.
- Lowercasing, punctuation/number removal.
- Custom stopword filtering.
- Lemmatization.
- Label encoding: 0 = fake, 1 = real.

Data Dictionary

Column	Description
title	Article headline

Column	Description
text	Article body
subject	News topic/category
date	Publication date
label	0 = Fake, 1 = Real
clean_text	Preprocessed full text

Methods

- **Exploratory Data Analysis (EDA):**
  Examined label balance, word distributions, and topic themes.
- **Classical ML:**
  TF-IDF vectorization, Logistic Regression, Random Forest.
  Model evaluation: accuracy, ROC-AUC, confusion matrix, F1-score.
- **Deep Learning:**
  Fine-tuned BERT for text classification.
  Used Hugging Face Transformers library.
- **Explainability:**
  SHAP analysis for BERT.

Analysis

Exploratory Data Analysis

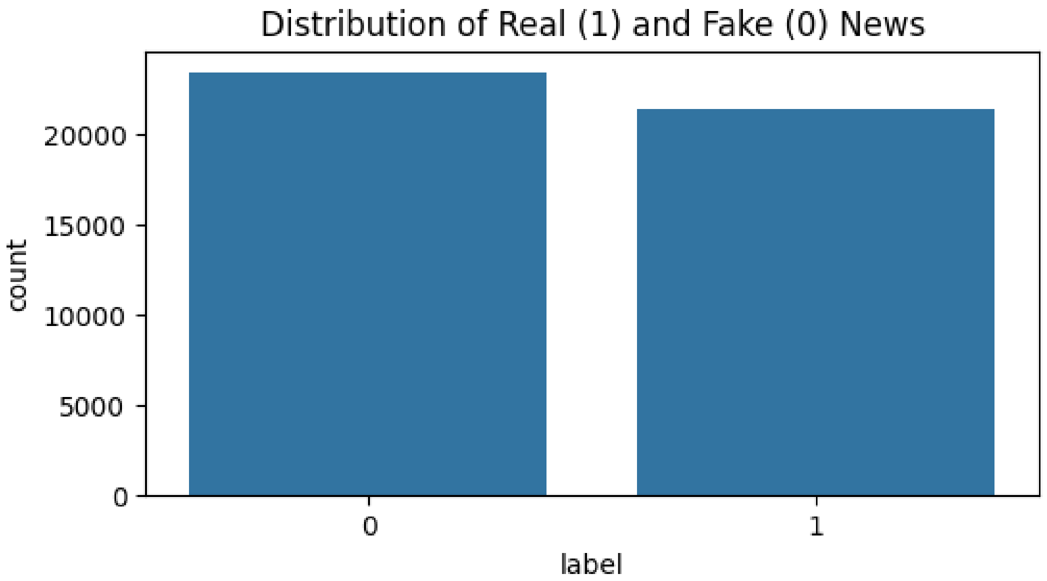

Figure 1. *Label distribution (bar plot) showing fake/real news counts.*

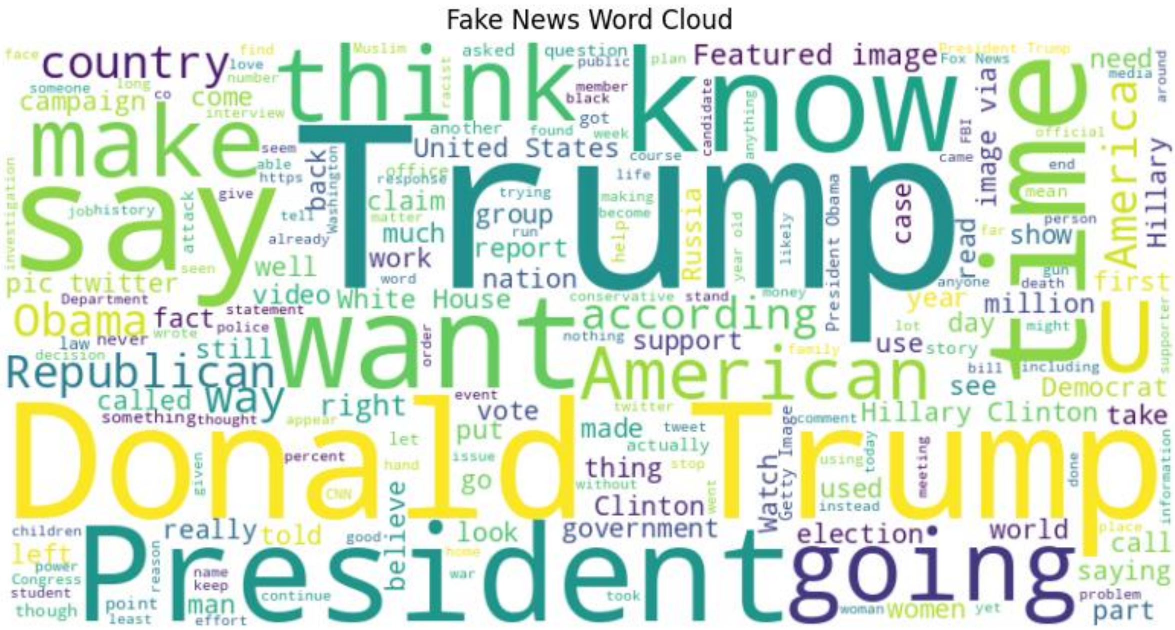

Figure 2. *Word cloud for Fake News (top words after cleaning).*

Figure 3. *Word cloud for True News (top words after cleaning).*

Findings:

- Dataset is moderately balanced.
- Fake news uses more viral or emotive words ("video", "image", "via", "breaking"), real news highlights agency names and neutral reporting.

Classical Model Results

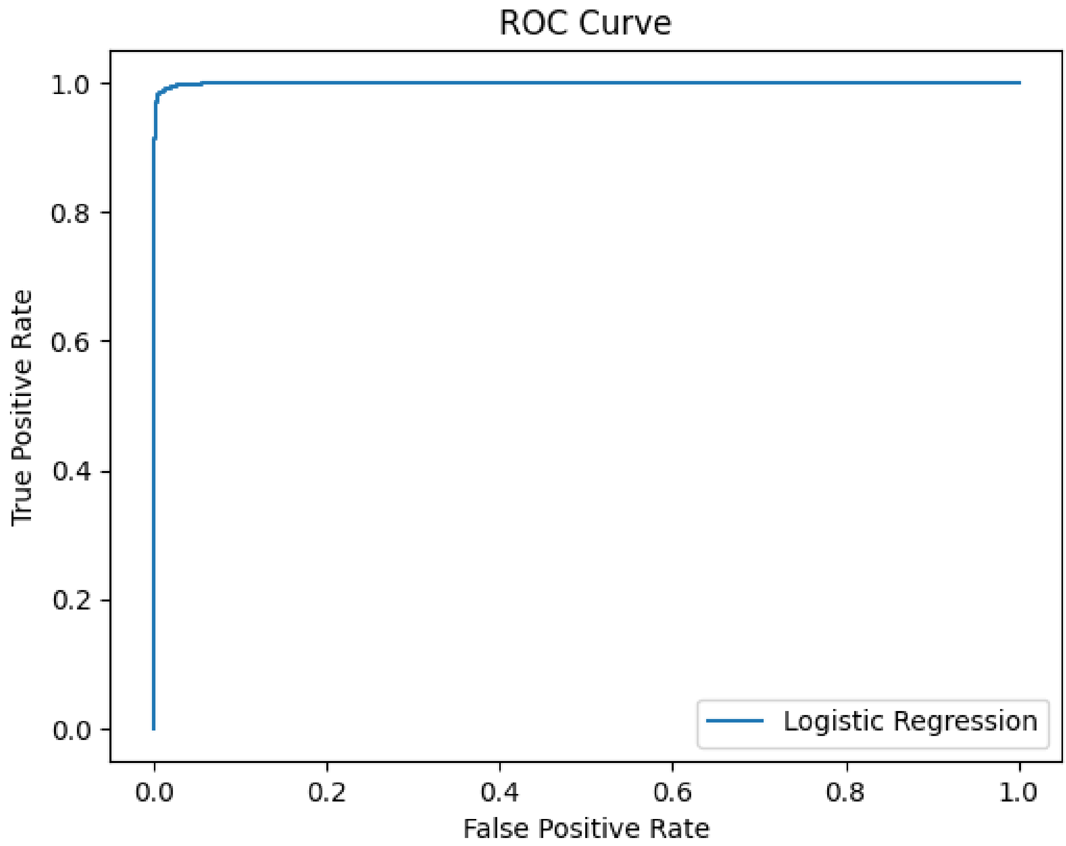

Figure 4. *ROC curve for Logistic Regression model.*

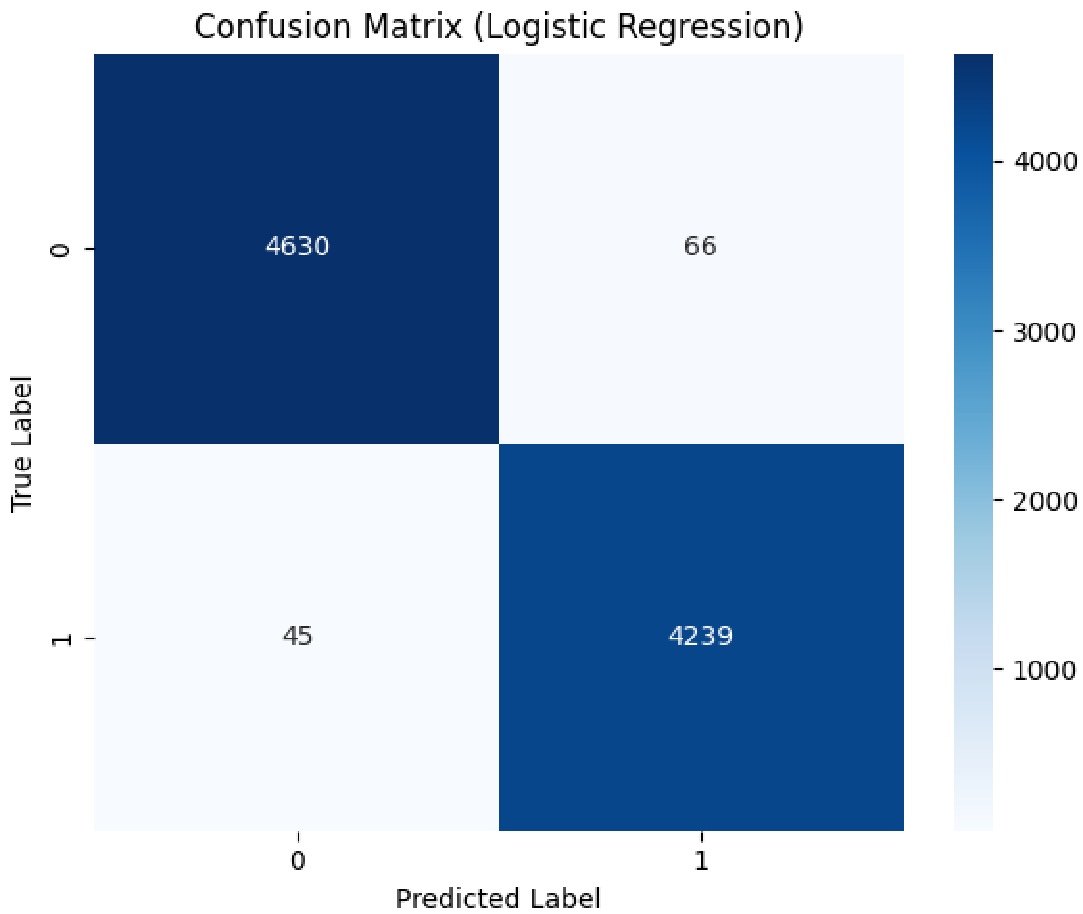

Figure 5. *Confusion Matrix for Logistic Regression.*

Performance:

- **Logistic Regression:** ~97% accuracy, AUC >0.97.
- **Random Forest:** Similar performance.
- **Feature Importance:**
  - Top real news predictors: "reuters", "washington", "president donald", "statement".
  - Top fake news predictors: "video", "image", "via", "breaking", "hillary".

<u>Top words predicting REAL news:</u>

1. reuters: 23.589
2. washington reuters: 10.014
3. wednesday: 6.176
4. president donald: 5.816
5. tuesday: 5.770
6. washington: 5.440
7. thursday: 5.297
8. friday: 4.993
9. reuters president: 4.794
10. monday: 4.781

Top words predicting FAKE news:

1. video: -10.269
2. image: -8.813
3. via: -8.799
4. gop: -6.093
5. president trump: -6.091
6. hillary: -6.040
7. image via: -5.160
8. obama: -4.930
9. america: -4.494
10. american: -4.481

Deep Learning (BERT) Results

- **BERT** trained on a subset due to computational constraints.
- Achieved similarly high accuracy and ROC-AUC as classical models.

[800/800 2:51:44, Epoch 2/2]

Epoch	Training Loss	Validation Loss
1	0.018200	0.046048
2	0.044800	0.027616

[100/100 05:46]

```
{'eval_loss': 0.027616212144494057,
 'eval_runtime': 349.4538,
 'eval_samples_per_second': 2.289,
 'eval_steps_per_second': 0.286,
 'epoch': 2.0}
```

- **BERT Explainability:**
  Used SHAP to visualize token importance for predictions.

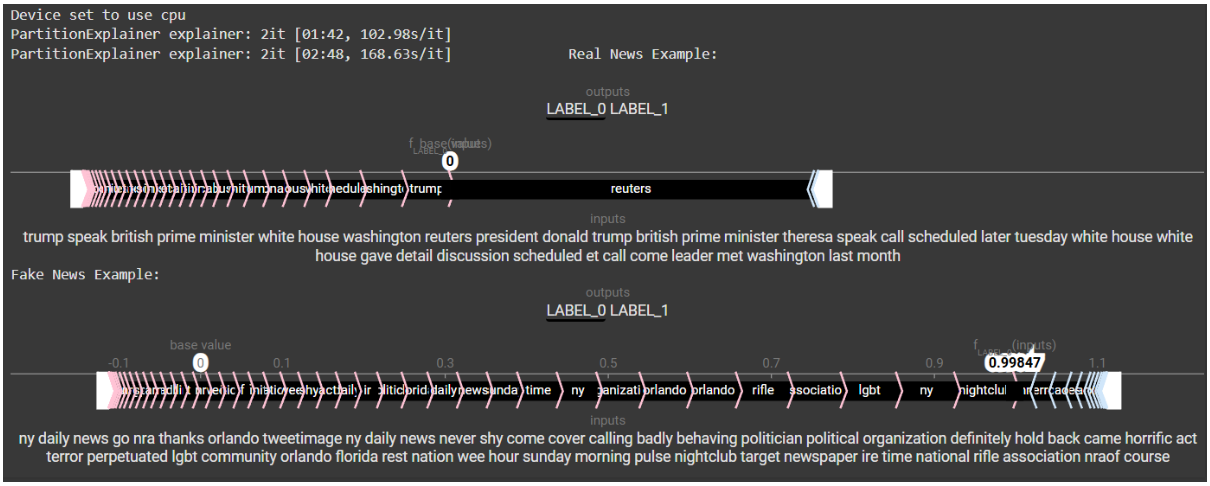

Figure 6. *SHAP explanation for one BERT prediction (screenshot/token plot).*

Conclusion

Automated models can reliably distinguish fake from real news on benchmark datasets. Logistic Regression is highly effective with TF-IDF, while BERT matches or slightly exceeds performance, especially for more complex or future tasks. SHAP explainability builds trust by clarifying model decisions.

Assumptions

- Labels are accurate and generalize to new data.
- Preprocessing preserves key meaning.
- Evaluation metrics reflect real-world deployment.

Limitations

- Dataset may contain bias or artifacts.
- Real-world fake news is more diverse than this dataset.
- BERT’s 512-token limit restricts very long article processing.
- SHAP can be slow for deep learning.

Challenges

- Generalization to new types of fake news.
- Computational cost for BERT and SHAP.
- Explaining decisions to non-technical stakeholders.

Future Uses / Additional Applications

- Browser extension, content moderation, or social media plugin.
- Cross-lingual and multimodal (image/video/text) fake news detection.
- Real-time flagging and user feedback systems.

Recommendations

- Routinely retrain with new data to prevent model drift.
- Monitor for unintended bias.
- Pair automation with human review in high-stakes scenarios.
- Use explainability tools (e.g., SHAP) for model audits.

Implementation Plan

1. **Data Cleaning:** Prepare and preprocess all input.
2. **Baseline Modeling:** Classical ML for fast, explainable results.
3. **Deep Learning:** Fine-tune BERT for text.
4. **Evaluation:** Use metrics, confusion matrix, ROC, SHAP.
5. **Deployment (optional):** API/web service.
6. **Reporting:** Document results, visuals, and findings.

Ethical Assessment

- Guard against bias (e.g., not unfairly flagging certain topics or sources).
- Preserve user privacy (no personal data collected).
- Ensure transparency in predictions (explainability via SHAP).
- Use human oversight for critical use-cases.
- Disclose limitations—model is not a replacement for expert review.

APA References

Bisaillon, C. (2019). *Fake and Real News Dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference* *of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Wang, W. Y. (2017). "Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 422–426. https://doi.org/10.18653/v1/P17-2067

Zhang, Z., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025. https://doi.org/10.1016/j.ipm.2019.102025

---

10 Audience Questions (For Milestone 4)

1. How do you define and label fake news in this study?
2. What are the main linguistic features or words that distinguish fake from real news?
3. How do classical ML and BERT compare in terms of performance and transparency?
4. What are the limitations or risks of deploying this model in production?
5. How do you address the possibility of model bias or unfairness?
6. Can this model be fooled or adversarially attacked?
7. How will the model handle newly emerging fake news tactics or topics?
8. Is it possible to adapt this approach to other languages or types of misinformation?
9. How transparent and explainable are the predictions to a non-technical end user?
10. What are the most important next steps to improve and deploy the system?